MST121 EBD

The Open University

USING MATHEMATICS

Exercise Booklet D

Exercise Booklet

D

SUP 86873 8
2.1

Contents

Exercises for Chapter D1

Section 1

Exercise 1.1

Consider a pack of fifty cards, numbered from 1 to 50, which is completely shuffled. A card is drawn from the pack. What is the probability that the number on this card

(a) is 17;

(b) is not 17;

(c) is divisible by 5;

(d) ends in the digit 3;

(e) is 51?

Exercise 1.2

In the UK National Lottery, described in Chapter D1, Activity 1.4, there are 13 983 816 different selections of six numbers between 1 and 49 (inclusive).

(a) What is the probability that a person will win a share of the jackpot if he or she buys 1000 tickets, all with different selections?

(b) How many tickets would have to be bought to give a player an even chance (probability $\frac{1}{2}$) of winning a share of the jackpot?

Exercise 1.3

A sock drawer in a very dark room contains 50 orange and 50 pink socks, each identical to the touch. How many socks must be taken out of the drawer to be sure of having taken a pair of the same colour?

Section 3

Exercise 3.1

(a) A bag contains 2 red balls, 5 yellow balls and 6 green balls, each identical to the touch. If a ball is taken from the bag without looking, what is the probability that the colour of the ball is red?

(b) If a letter is chosen at random from the word 'petunia', what is the probability that the letter is a vowel?

Exercise 3.2

Assume that a human population contains equal numbers of boys and girls.

(a) What is the probability that, in a randomly selected family with two children, the children are both girls?

(b) What is the probability that, in a randomly selected family with two children where the elder child is a girl, the children are both girls?

(c) What is the probability that, in a randomly selected family with two children where one child is a girl, the children are both girls?

Exercise 3.3

Andrew and Morag are equally skilled archers. In a particular practice session, Andrew has two arrows to shoot and Morag has one. The winning arrow is the one that lands closest to the bull's-eye. Which one of the following statements is true? (Assume that it is not possible for two arrows to land equally close to the bull's eye.)

(a) There are three arrows in total and Andrew has two of them. The probability that he will win is therefore $\frac{2}{3}$.

(b) There are four possible outcomes: both of Andrew's arrows are closer to the bull's-eye than is Morag's arrow; Andrew's first arrow (only) is closer; Andrew's second arrow (only) is closer; both of Andrew's arrows are further away. Only in the last case does Andrew lose, so his probability of winning is $\frac{3}{4}$.

Exercise 3.4

Two people each choose at random an integer between 1 and 10, inclusive.

(a) Find the probability that the two people do not both choose the number 3.

(b) This process is now repeated twenty times.

(i) Find the probability that the two people do not both choose the number 3 on any of the twenty occasions.

(ii) Find the probability that they both choose the number 3 on at least one of the twenty occasions.

Exercise 3.5

A regular octahedral die has eight faces, labelled $1, 2, \ldots, 8$. It is equally likely to land with any one of these eight faces uppermost, and the score obtained is the number that appears on the uppermost face. Three such octahedral dice are rolled together repeatedly.

(a) Find the probability that, in a single roll of the three dice, the scores obtained are the same.

(b) Find the probability that a 'triple 5' is obtained in a single roll of the three dice.

(c) Find the probability that there is no 'triple 5' in the first 10 rolls of the three dice.

(d) Find the probability that there is at least one 'triple 5' in the first 10 rolls of the three dice.

Exercise 3.6

Pete, Nick and Ian each choose at random an integer between 1 and 10, inclusive.

(a) Find the probability that Pete, Nick and Ian all choose different numbers.

(b) Find the probability that at least two of them choose the same number.

(c) Find the probability that they all choose the same number.

(d) Find the probability that exactly two of them choose the same number.

(e) Find the probability that Nick's chosen number is greater than Ian's chosen number.

Exercise 3.7

Three pieces of toast, buttered on one side, are dropped together. The probability of any one piece landing butter-side down is 0.62.

(a) Find the probability that no piece of toast lands butter-side down.

(b) Find the probability that at least one piece of toast lands butter-side down.

(c) Find the probability that exactly two pieces of toast land butter-side down.

(d) Find the probability that at least two pieces of toast land butter-side down.

Exercise 3.8

Three friends visit the same gym as regularly as they can. The first goes to the gym, on average, one evening in every two, the second goes one evening in every three, and the third goes one evening in every five. In each case, no evening is more likely to be selected than any other.

(a) What is the probability that at least one of the friends is present at the gym on any given evening?

(b) What is the probability that at least two of them are present on any given evening?

Exercise 3.9

A coin is biased so that when it is tossed, the outcome is four times more likely to be a tail than a head. What are the probabilities of obtaining a head and of obtaining a tail for this coin?

Section 4

Exercise 4.1

Consider a multiple choice assessment paper, on which each question has five possible answers. In each case, just one of these answers is correct. A student who is taking this paper decides to answer the questions in order but to guess the answers at random.

(a) Find the probability that the answer given to Question 1 is correct.

(b) Find the probability that Question 1 is answered incorrectly whereas the answer to Question 2 is correct.

(c) Find the probability that the first correct answer given is for Question 4.

(d) What is the probability that Questions 1–8 are all answered incorrectly?

(e) What is the probability that at least one of Questions 1–8 is answered correctly?

(f) Approximately how many guesses, on average, would have to be made in order to obtain a correct answer?

(g) If there are 25 questions in total on the paper, what is the probability of giving incorrect answers to all of them?

Exercise 4.2

For the situation described in Exercise 3.5, suppose that the three octahedral dice are rolled until a 'triple 5' is obtained. How many times would you expect the dice to be rolled; that is, what is the mean number of rolls required to obtain a 'triple 5'?

Exercise 4.3

A local delicatessen makes up a hamper of luxury goods each week, and sells raffle tickets to customers with the hamper as prize and any profits donated to charity. It is estimated that the probability of winning the hamper on each occasion is $\frac{1}{150}$ for each ticket.

(a) A customer buys one ticket each week. Show that the probability that she wins the hamper at least once within two years is just greater than 0.5.

(b) How many tickets would the customer have to buy in any single week in order to have a probability of at least 0.5 of winning the hamper?

(c) If the customer buys one ticket per week, what is the average number of weeks between successive occasions on which she wins the hamper?

Section 5

Exercise 5.1

A supermarket chain includes a free toy dinosaur in each of their sealed lunch boxes for children. There are seven different types of dinosaur to collect. How many lunch boxes would have to be purchased, on average, in order to collect at least one of each type of dinosaur? What assumptions have you made?

Exercise 5.2

Four MST121 students go to a restaurant for a meal after their last tutorial. There are eight main courses on the menu, and each dish is equally palatable to each of the students.

(a) What is the probability that two particular students order the same main course?

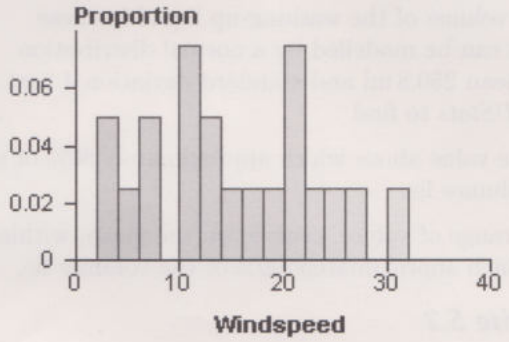
(b) What is the probability that at least two of the four students select the same main course?

Exercises for Chapter D2

Section 1

Exercise 1.1

The histogram below arises from data of daily average wind speeds in Alaska. (The source does not report the units in which the wind speed was measured.)



(a) Is a normal distribution a suitable model for variation in this data?

(b) On a separate diagram, but using a similar scale, sketch a curve which might be a suitable model for the variation in the data.

(c) Explain, with reference to the sketch in part (b), how this model could be used to estimate the proportion of daily average wind speeds of more than 20 units in magnitude.

Section 2

Exercise 2.1

The marks out of 10 for six children in a particular class for an arithmetic test were as follows.

7 5 4 5 6 9

Find, by hand,

(a) the sample mean;

(b) the sample standard deviation.

Exercise 2.2

Set A below is a sample of marks obtained in a spelling test from a particular class of children, and set B is a sample of marks obtained from another class in the same test.

Set A:	9	10	11	12	13
Set B:	1	6	11	16	21

(a) Calculate the following for each of sets A and B:

(i) the sample mean;

(ii) the sample standard deviation.

(b) What can you deduce from your answers to part (a)?

Exercise 2.3

Which of the following are true?

(a) The standard deviation of a sample of data measures the variability of the data about the sample mean. All of the data are involved in its calculation.

(b) The standard deviation of a set of numbers measures the mean distance of the numbers from the mean.

(c) The standard deviation of a data set has the same units as the data. For example, the standard deviation of a set of heights in cm will also be in cm.

(d) Comparing the standard deviations of different data sets is a way of judging the relative spreads of those data sets.

(e) A standard deviation cannot have a negative value.

Section 3

The exercises in this section require the use of *OUStats*.

Exercise 3.1

The file IQ.OUS gives the results of an intelligence test of 112 children. The IQ scores are measured to the nearest integer, using the Stanford revision of the Simon–Binet intelligence scale.

(a) What is the least value of IQ recorded?

(b) What first interval starting value and interval width would it be sensible to choose in order to obtain a frequency diagram of the data?

(c) For the data provided, produce

(i) a frequency diagram;

(ii) a histogram.

(d) What is the difference between the two diagrams in part (c)?

Exercise 3.2

The file BEARS.OUS contains data on the average wind speed on each of 20 days in Alaska. (The source does not report the units in which the wind speed was measured.) Plot a histogram of the data, after making appropriate choices for the first interval starting value and interval width.

Section 4

Exercise 4.1 requires the use of *OUStats*. The data given in this exercise need to be entered into a new file. Details of how to do this are given in the Appendix of Computer Book D. However, it is not an expectation of the course that you should be able to enter data, so you may prefer to omit this exercise.

Exercise 4.1

On Tuesday 18 March 2003, *The Scotsman* published the number of hours of sunshine on the previous day for a number of places in Scotland, as shown in the table below.

Place	Hours of sunshine
Aberdeen	5.3
Aviemore	9.0
Edinburgh	0
Eskdalemuir	10.0
Glasgow	9.8
Kinloss	6.7
Tiree	9.7

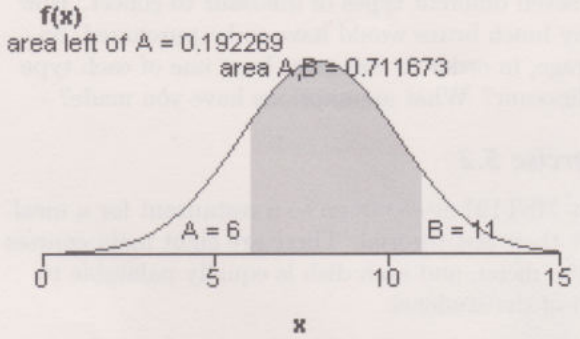
Use *OUStats* to find, for this set of data,

(a) the sample mean;

(b) the sample standard deviation.

Exercise 4.2

The *OUStats* diagram below shows a normal curve with mean 8 and standard deviation 2.3. Suppose that this normal curve is a model for the variability of some attribute within a large population.



The diagram gives values for the area under the curve to the left of A and for the area between A and B. Using this information, and working to 3 decimal places, estimate the proportion of the population for whom the value of the attribute is

(a) more than 6;

(b) less than 11;

(c) between 8 and 11.

Section 5

The exercises in this section require the use of *OUStats*.

Exercise 5.1

A manufacturer of washing-up liquid produces a liquid with a new fragrance, in bottles labelled to indicate a volume of 250 ml. The variation in the actual volume of the washing-up liquid in these bottles can be modelled by a normal distribution with mean 250.8 ml and standard deviation 4.3 ml. Use *OUStats* to find

(a) the value above which approximately 80% of the volumes lie;

(b) a range of values, centred on the mean, within which approximately 90% of the volumes lie.

Exercise 5.2

Variation in the time for a visit to a particular website can be modelled by a normal distribution with mean 32.5 minutes and standard deviation 7.5 minutes.

(a) What is the time within which approximately 50% of visits are completed?

(b) What is the time within which approximately 90% of visits are completed?

(c) Find a range of values for the time, centred on the mean, within which approximately 99% of visits are completed.

Exercise 5.3

State the proportion of values in a normal distribution that are more than k standard deviations from the mean, for each of the following values of k.

(a) 1 (b) 1.5 (c) 2.25 (d) 2.5 (e) 3

Section 6

Exercise 6.1

The weights in kg of adult males in a particular town are modelled by a normal distribution with mean $\mu = 72.4$ and standard deviation $\sigma = 14.7$. For each of the following proportions of adult males in the town, find a range within which the weights of that proportion should lie, according to the model.

(a) 99.7% (b) 90% (c) 95% (d) 99%

Exercises for Chapter D3

Section 1

Exercise 1.1

The distribution of volumes of the contents of bottles of vinegar, labelled as containing 250 ml, has mean 251 ml and standard deviation 2.5 ml.

(a) Find the mean and standard deviation of the sampling distribution of the mean, for samples of 60 bottles.

(b) Find a range of values within which the mean volumes of approximately 95% of samples of size 60 will lie.

Exercise 1.2

The distribution of the weights of contents of jars of chocolate spread, labelled as containing 400 g, has mean 401.6 g and standard deviation 4.2 g.

(a) What is the standard error of the mean (to 3 decimal places), for samples of 45 jars?

(b) Find a range of values within which the mean weights of approximately 95% of samples of size 45 will lie.

Section 2

Exercise 2.1

(a) By law, the mean volume of liquid contained in cans of a particular soft drink should be at least the nominal volume of 330 ml. The mean volume of liquid in a sample of 50 cans is 331.7 ml, and the sample standard deviation is 5.75 ml.

(i) Calculate a 95% confidence interval for the mean volume of liquid in cans of the soft drink.

(ii) What can the soft drink manufacturer conclude from this confidence interval about the mean volume of liquid in cans?

(b) The production manager decides to take another sample of cans and to calculate a second 95% confidence interval for the mean volume using this sample. She wants the width of this confidence interval to be half the width of the confidence interval calculated in part (a)(i).

(i) Approximately what sample size should be taken?

(ii) Will the width of the 95% confidence interval calculated from a sample of this size be exactly one half of the width of the confidence interval in part (a)(i)?

Exercise 2.2

The number of words was noted in each of 30 sentences chosen at random from a popular novel. The mean length of sentences in the sample was 19.0 words, and the standard deviation was 10.9 words.

(a) Calculate an approximate 95% confidence interval for the mean length of sentences in the novel.

(b) What does this confidence interval tell you?

Exercise 2.3

(a) Past experience has shown that the lifetimes for a particular type of battery have mean 50 hours and standard deviation 7.5 hours.

(i) Find the mean and standard deviation of the sampling distribution of the mean, for samples of size 100 from this population.

(ii) Find a range of values within which the mean lifetimes of approximately 95% of samples of 100 batteries will lie.

(iii) How does this range differ from a 95% confidence interval?

(b) New operational procedures are introduced by the battery manufacturer. To investigate the effect of the changes, the lifetimes of a sample of 100 batteries are recorded. The sample mean is 49 hours and the sample standard deviation is 4.8 hours.

(i) Calculate an approximate 95% confidence interval for the mean lifetime of batteries produced under the new procedures.

(ii) What should the manufacturer conclude about the effectiveness of the new procedures, as compared with the old?

Exercise 2.4

Data were collected in the 1890s on the heights of 1079 father–mother pairs. These data are in the file COUPLES.OUS, from which it can be ascertained that the differences between father's height and mother's height in the sample have mean 5.15 inches and standard deviation 3.03 inches.

(a) Calculate a 95% confidence interval for the mean difference in height between a father and mother in the 1890s. What does this confidence interval tell you about the difference in height between fathers and mothers in the population from which this sample was taken?

(b) Is it likely that the mean difference in height between husbands and wives was 4 inches or less in this population, and that the difference observed in the sample of 1079 couples was just due to chance?

Section 3

Exercise 3.1 requires the use of *OUStats*.

Exercise 3.1

The file WEASEL.OUS contains two samples of data on the lengths of male weasels, taken from Sussex and from Northumberland respectively.

(a) Calculate 95% confidence intervals for the mean lengths of male weasels from Sussex and from Northumberland.

(b) Using the values of sample size, sample mean and sample standard deviation provided by *OUStats*, calculate the same confidence interval for Northumberland weasels by hand, and check that your answer matches that from part (a).

Exercises for Chapter D4

Section 1

Exercise 1.1

The lengths, in seconds, of the tracks on a music CD are as follows.

132 211 154 150 150
153 179 175 155 212

(a) Find the median, lower quartile, upper quartile, range and interquartile range for this sample.

(b) Draw a boxplot to represent these data.

Exercise 1.2

Chris and Emily record their scores for 15 rounds of golf. Chris's scores are as follows.

78 73 81 80 68 74 74 71
71 71 81 75 73 80 84

Emily's scores are as follows.

80 70 74 77 77 73 71 72
83 71 76 73 79 84 83

(a) Find the median, lower quartile, upper quartile, range and interquartile range for Chris's scores and for Emily's scores.

(b) Construct boxplots on a common axis for each of their scores.

(c) Chris claims that he is a better golfer than Emily. Use your answer to part (b) as evidence either to support or refute this claim.

Section 2

Exercise 2.1 requires the use of *OUStats*.

Exercise 2.1

The file WEASEL.OUS contains two samples of data on the lengths of male weasels, taken from Sussex and from Northumberland respectively. (These data were referred to in Exercise 3.1 for Chapter D3.)

(a) Obtain boxplots of the two samples of data on a single diagram.

(b) What do these boxplots tell you about the lengths of male weasels in Sussex as compared with those in Northumberland?

Section 3

Exercise 3.1

The number of words was noted in each of 70 sentences selected at random from two romantic novels written by different authors; 35 sentences were selected from each book. A summary of the data is given below.

		Words per sentence	
	Sample size	Sample mean	Sample standard deviation
Author A	35	22.3	17.2
Author B	35	21.4	13.4

The two-sample z-test is to be used to investigate whether there is a difference between the mean sentence lengths in the book by Author A and the mean sentence lengths in the book by Author B. The null and alternative hypotheses may be written as

$$H_0 : \mu_A = \mu_B,$$
$$H_1 : \mu_A \neq \mu_B,$$

where μ_A is the mean sentence length in the book by Author A, and μ_B is the mean sentence length in the book by Author B.

(a) Find the estimated standard error of the difference between the two sample means, correct to 2 decimal places.

(b) Calculate the value of the test statistic, z.

(c) Use the two-sample z-test to investigate whether there is a difference between the mean sentence lengths in the book by Author A and the mean sentence lengths in the book by Author B.

Section 4

Exercise 4.1 requires the use of *OUStats*.

Exercise 4.1

The file WEASEL.OUS contains two samples of data on the lengths of male weasels, taken from Sussex and from Northumberland respectively. (These data were referred to in Exercise 3.1 for Chapter D3 and in Exercise 2.1.)

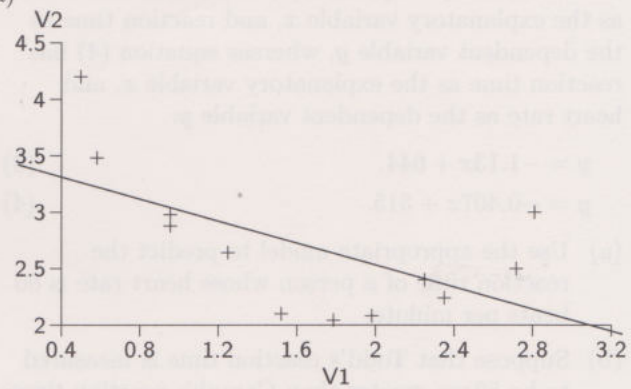
Use the two-sample z-test to investigate whether there is a difference between the mean lengths of male weasels from Sussex and male weasels from Northumberland.

Section 5

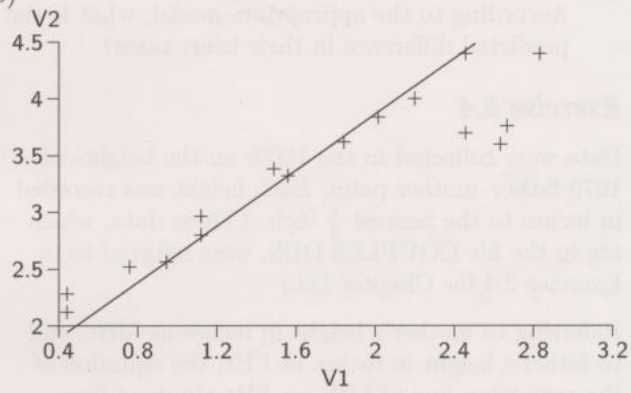
Exercise 5.1

Explain, in each case below, why the line shown does not fit the data well.

(a)



(b)



Exercise 5.2

Data were obtained on the rate at which a cricket chirps (measured in number of chirps per second) and the corresponding ambient temperature (in degrees Fahrenheit) in the vicinity of the cricket. Regression lines were fitted to these data.

With ambient temperature as the explanatory variable x, and chirp rate as the dependent variable y, the corresponding regression line has equation

$$y = 0.2119x - 0.3091. \qquad (1)$$

With chirp rate as the explanatory variable x, and ambient temperature as the dependent variable y, the corresponding regression line has equation

$$y = 3.291x + 25.23. \qquad (2)$$

Use the appropriate model to answer each of the following. (You may assume in each case that the explanatory variable values concerned lie within the ranges for which the data were collected.)

(a) Find the predicted ambient temperature when the cricket chirp rate is 18.0 chirps per second.

(b) If the ambient temperature drops by 5°F, find the corresponding difference predicted for the cricket chirp rate.

Exercise 5.3

It is thought that reaction time (in milliseconds) is related to heart rate (in beats per minute). Straight-line models, derived as regression lines from data collected in a hospital, can be used to describe this relationship. Equation (3) below has heart rate as the explanatory variable x, and reaction time as the dependent variable y, whereas equation (4) has reaction time as the explanatory variable x, and heart rate as the dependent variable y:

$$y = -1.13x + 644, \quad (3)$$

$$y = -0.407x + 315. \quad (4)$$

(a) Use the appropriate model to predict the reaction time of a person whose heart rate is 80 beats per minute.

(b) Suppose that Todd's reaction time is measured to be 50 ms greater than Caragh's reaction time. According to the appropriate model, what is the predicted difference in their heart rates?

Exercise 5.4

Data were collected in the 1890s on the heights of 1079 father–mother pairs. Each height was recorded in inches to the nearest $\frac{1}{4}$ inch. (These data, which are in the file COUPLES.OUS, were referred to in Exercise 2.4 for Chapter D3.)

Referring to mother's height in inches as MHt, and to father's height in inches as FHt, the equation of the regression line of MHt on FHt obtained from these data is

$$\text{MHt} = 44.98 + 0.2595\,\text{FHt}.$$

(a) Use the above equation to estimate in each case the mean height of mothers for whom the corresponding fathers have the following heights, in inches.

(i) 62 (ii) 66 (iii) 70 (iv) 73

(b) What do these results tell you about the relationship between fathers' heights and corresponding mothers' heights in the population from which this sample was drawn in the 1890s?

Section 6

Exercise 6.1 requires the use of *OUStats*. The data given in this exercise need to be entered into a new file. Details of how to do this are given in the Appendix of Computer Book D. However, it is not an expectation of the course that you should be able to enter data, so you may prefer to omit this exercise.

Exercise 6.1

The following table gives data on the rate at which a cricket chirps and the corresponding ambient temperature in the vicinity of the cricket.

Ambient temperature (°F)	Chirps per second
88.6	20.0
71.6	16.0
93.3	19.8
84.3	18.4
80.6	17.1
75.2	15.5
69.7	14.7
82.0	17.1
69.4	15.4
83.3	16.2
79.6	15.0
82.6	17.2
80.6	16.0
83.5	17.0
76.3	14.4

(a) Obtain a scatterplot to illustrate these data, with ambient temperature along the x-axis and chirp rate along the y-axis. Add the regression line to the scatterplot.

(b) Is there any evidence of a relationship between the ambient temperature and the chirp rate of crickets?

(c) Find the equation of the regression line of y on x for these data, in each of the following cases.

(i) Take x to be the ambient temperature and y to be the chirp rate (as in part (a) above).

(ii) Take x to be the chirp rate and y to be the ambient temperature.

Solutions for Chapter D1

Solution 1.1

(a) $P(\text{number is } 17) = \frac{1}{50}$

(b) $P(\text{number is not } 17) = \frac{49}{50}$

(c) $P(\text{number is divisible by } 5) = \frac{10}{50} = \frac{1}{5}$

(d) $P(\text{number ends in the digit } 3) = \frac{5}{50} = \frac{1}{10}$

(e) $P(\text{number is } 51) = 0$

Solution 1.2

(a) Each of the 13 983 816 selections is equally likely to occur. The probability of winning a share of the jackpot with 1000 different selections is

$$\frac{1000}{13\,983\,816} \simeq 0.000\,07.$$

(b) We seek the number n of tickets such that

$$\frac{n}{13\,983\,816} = \frac{1}{2}.$$

The solution of this equation is $n = 6\,991\,908$, so nearly seven million tickets would have to be bought in order to provide an even chance of winning a share of the jackpot.

Solution 1.3

The answer is three. (The answer is not 51, which is a common immediate response, and which is the number of socks that must be taken to be sure of having a pair *not* of the same colour.)

The first sock is either orange or pink. The second sock could be different from the first (otherwise two would suffice). In this case, the third sock is the same colour as either the first sock or the second sock. So three socks must be taken from the drawer to ensure obtaining a pair of the same colour.

Solution 3.1

(a) There are 2 red balls out of $2 + 5 + 6 = 13$ balls in total. So, by rule (3.1), we have $P(\text{red}) = \frac{2}{13}$.

(b) The letters 'e', 'u', 'i' and 'a' are vowels, while 'p', 't' and 'n' are consonants. Hence there are 4 vowels in the total of 7 letters, so by rule (3.1), the required probability is $P(\text{vowel}) = \frac{4}{7}$.

Solution 3.2

We use B to represent a boy and G to represent a girl, with the elder child placed first in order.

(a) The equally-likely outcomes are BB, BG, GB, GG, so

$$P(\text{both girls}) = \tfrac{1}{4}.$$

(b) The equally-likely outcomes are GB, GG, so

$$P(\text{both girls}) = \tfrac{1}{2}.$$

This answer can also be obtained by thinking just about the younger child, for which the equally-likely outcomes are B and G.

(c) The equally-likely outcomes are BG, GB, GG, so

$$P(\text{both girls}) = \tfrac{1}{3}.$$

Solution 3.3

To answer such a question, it is necessary to identify the equally-likely outcomes. The two archers are equally skilled (and each of Andrew's two arrows will be dispatched with equal skill), so there are six equally likely arrangements of the three arrows, as described from nearest to furthest from the bull's-eye. If Andrew's two arrows are denoted by A_1, A_2 and Morag's arrow by M, then these arrangements are A_1A_2M, A_1MA_2, A_2A_1M, A_2MA_1, MA_1A_2 and MA_2A_1. In four of these cases, one of Andrew's arrows is closest to the bull's-eye, so the probability that he will win is $\frac{4}{6} = \frac{2}{3}$. Hence statement (a) is true. By contrast, the four possible outcomes described in statement (b) are not equally likely.

Solution 3.4

(a) For each person, we have

$$P(\text{chooses 3}) = \tfrac{1}{10},$$

and hence (by rule (3.3))

$$P(\text{both choose 3}) = \tfrac{1}{10} \times \tfrac{1}{10} = \tfrac{1}{100}.$$

It follows from rule (3.2) that

$$P(\text{not both choose 3}) = 1 - \tfrac{1}{100} = \tfrac{99}{100}.$$

(b) (i) Using rule (3.3) once more gives

$$P(\text{never both choose 3}) = \left(\tfrac{99}{100}\right)^{20} = 0.8179 \quad \text{(to 4 d.p.)}.$$

(ii) By rule (3.2), we have

$$P(\text{both choose 3 at least once}) = 1 - P(\text{never both choose 3}) = 1 - \left(\tfrac{99}{100}\right)^{20} = 0.1821 \quad \text{(to 4 d.p.)}.$$

(This is similar to the approach used to solve De Méré's problem in Chapter D1, Activity 3.12.)

Solution 3.5

(a) The first die can land with any one of the eight faces uppermost. The second and third dice then each have probability $\frac{1}{8}$ of landing with the same face uppermost as the first die so, using rule (3.3), we have

$$P(\text{same score on three dice}) = \tfrac{1}{8} \times \tfrac{1}{8} = \tfrac{1}{64}.$$

(Alternatively, there are $8 \times 8 \times 8$ possible outcomes for the three dice, and just 8 of these outcomes correspond to all scores being the same, so the probability required is $\frac{8}{8\times8\times8} = \frac{1}{64}$.)

(b) The probability that the score on a single die is 5 is $\frac{1}{8}$ so, by rule (3.3), we have

$$\begin{aligned} P(\text{triple 5}) &= \tfrac{1}{8} \times \tfrac{1}{8} \times \tfrac{1}{8} \\ &= \tfrac{1}{512} = 0.0020 \quad \text{(to 4 d.p.)}. \end{aligned}$$

(Alternatively, a 'triple 5' is just one of the $8 \times 8 \times 8$ possible outcomes for the three dice, leading to the same result.)

(c) For a single roll, using rule (3.2), we have

$$\begin{aligned} &P(\text{not a triple 5}) = 1 - P(\text{triple 5}) \\ &= 1 - \tfrac{1}{512} = \tfrac{511}{512} = 0.9980 \quad \text{(to 4 d.p.)}. \end{aligned}$$

It follows from rule (3.3) that

$$\begin{aligned} &P(\text{no triple 5 in 10 rolls}) \\ &= \left(\tfrac{511}{512}\right)^{10} = 0.9806 \quad \text{(to 4 d.p.)}. \end{aligned}$$

(d) Using rule (3.2) again gives

$$\begin{aligned} &P(\text{at least one triple 5 in 10 rolls}) \\ &= 1 - P(\text{no triple 5 in 10 rolls}) \\ &= 1 - \left(\tfrac{511}{512}\right)^{10} = 0.0194 \quad \text{(to 4 d.p.)}. \end{aligned}$$

Solution 3.6

(a) Pete can choose any one of the 10 numbers. Nick can choose a different number from Pete in 9 out of 10 ways (hence with probability $\frac{9}{10}$), and Ian can choose a number different from both Pete and Nick in 8 out of 10 ways (that is, with probability $\frac{8}{10}$). Hence the probability that all three choose different numbers is

$$\tfrac{9}{10} \times \tfrac{8}{10} = \tfrac{72}{100}.$$

(b) We have

$$\begin{aligned} &P(\text{at least two choose same number}) \\ &= 1 - P(\text{all choose different numbers}) \\ &= 1 - \tfrac{72}{100} = \tfrac{28}{100}. \end{aligned}$$

(c) Pete can choose any one of the 10 numbers. The probability that either of Nick or Ian chooses the same number as Pete is $\frac{1}{10}$, so the probability that both do so is

$$\tfrac{1}{10} \times \tfrac{1}{10} = \tfrac{1}{100}.$$

(d) We have

$$\begin{aligned} &P(\text{exactly two choose same number}) \\ &= P(\text{at least two choose same number}) \\ &\quad - P(\text{all three choose same number}) \\ &= \tfrac{28}{100} - \tfrac{1}{100} = \tfrac{27}{100}, \quad \text{from parts (b) and (c)}. \end{aligned}$$

(e) The probability that Nick and Ian choose different numbers is $\frac{9}{10}$, as at the start of part (a). In half of these cases, the number chosen by Nick will be greater than that chosen by Ian, and vice versa in the other half of cases (by the symmetry of the set of outcomes involved). Hence the probability that Nick chooses a number greater than that chosen by Ian is

$$\tfrac{1}{2} \times \tfrac{9}{10} = \tfrac{9}{20}.$$

Solution 3.7

We use the notation D for a landing of one piece of toast butter-side down, and U for a landing butter-side up. Then DDU represents the outcome that the first and second pieces of toast land butter-side down while the third lands butter-side up, and similarly for all the other possible outcomes of dropping the three pieces of toast. We have

$$P(D) = 0.62 \quad \text{and} \quad P(U) = 1 - 0.62 = 0.38.$$

(a) The probability that no piece of toast lands butter-side down is

$$P(UUU) = (0.38)^3 = 0.0549 \quad \text{(to 4 d.p.)}.$$

(b) The probability that at least one piece lands butter-side down is

$$\begin{aligned} &P(\text{at least one down}) \\ &= 1 - P(UUU) \\ &= 1 - 0.0549 \quad \text{(from part (a))} \\ &= 0.9451 \quad \text{(to 4 d.p.)}. \end{aligned}$$

(c) The probability that exactly two pieces land butter-side down is

$$\begin{aligned} &P(DDU) + P(DUD) + P(UDD) \\ &= 3 \times 0.38 \times (0.62)^2 = 0.4382 \quad \text{(to 4 d.p)}. \end{aligned}$$

(d) The probability that at least two pieces land butter-side down is

$$\begin{aligned} &P(\text{exactly two down}) + P(\text{all three down}) \\ &= 3 \times 0.38 \times (0.62)^2 + (0.62)^3 \quad \text{(from part (c))} \\ &= 0.6765 \quad \text{(to 4 d.p.)}. \end{aligned}$$

Solution 3.8

(a) The probabilities that the friends are present individually at the gym on a given evening are, respectively, $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{5}$. The corresponding probabilities that they are not present are, respectively, $\frac{1}{2}$, $\frac{2}{3}$ and $\frac{4}{5}$. Hence the probability that none of them is present is

$$\tfrac{1}{2} \times \tfrac{2}{3} \times \tfrac{4}{5} = \tfrac{4}{15}.$$

We then have

$$\begin{aligned} &P(\text{at least one present}) \\ &= 1 - P(\text{none present}) \\ &= 1 - \tfrac{4}{15} = \tfrac{11}{15}. \end{aligned}$$

(b) The probability that at least two of the friends are present at the gym on a given evening is the sum of: (i) the probability that all three are present; (ii) the three different probabilities of one of them being absent while the other two are present. It follows that

$$\begin{aligned} &P(\text{at least two present}) \\ &= (\tfrac{1}{2} \times \tfrac{1}{3} \times \tfrac{1}{5}) + (\tfrac{1}{2} \times \tfrac{1}{3} \times \tfrac{4}{5}) \\ &\quad + (\tfrac{1}{2} \times \tfrac{2}{3} \times \tfrac{1}{5}) + (\tfrac{1}{2} \times \tfrac{1}{3} \times \tfrac{1}{5}) \\ &= \tfrac{1}{30} + \tfrac{4}{30} + \tfrac{2}{30} + \tfrac{1}{30} = \tfrac{8}{30} = \tfrac{4}{15}. \end{aligned}$$

Solution 3.9

If we put $p = P(\text{head})$, then we have $P(\text{tail}) = 4p$ from the information given. Since 'head' and 'tail' are the only possible outcomes of tossing the coin, we have

$$P(\text{head}) + P(\text{tail}) = 1 \quad \text{(by rule (3.2))}.$$

It follows that

$$p + 4p = 1, \quad \text{that is,} \quad 5p = 1,$$

so $p = \frac{1}{5}$. This gives

$$P(\text{head}) = \tfrac{1}{5} \quad \text{and} \quad P(\text{tail}) = \tfrac{4}{5}.$$

Solution 4.1

The probability that each answer is guessed correctly is $p = P(\text{success}) = \frac{1}{5}$. Let X be the number of answers required to obtain a first correct answer.

(a) The probability that the first answer is correct is

$$P(X = 1) = \tfrac{1}{5}.$$

(b) The probability that the first answer is incorrect but the second is correct is

$$P(X = 2) = (1 - p)p = \tfrac{4}{5} \times \tfrac{1}{5} = \tfrac{4}{25}.$$

(c) The probability that Question 4 yields the first correct answer is

$$\begin{aligned} &P(X = 4) = (1 - p)^3 p \\ &= \left(\tfrac{4}{5}\right)^3 \times \tfrac{1}{5} = \tfrac{64}{625} = 0.1024. \end{aligned}$$

(d) The probability of obtaining 8 successive incorrect answers is

$$\begin{aligned} &P(X > 8) = (1 - p)^8 \\ &= \left(\tfrac{4}{5}\right)^8 = \frac{65\,536}{390\,625} = 0.1678 \quad \text{(to 4 d.p.)}. \end{aligned}$$

(e) The probability that at least one of the first 8 answers is correct is

$$\begin{aligned} &P(X \leq 8) = 1 - P(X > 8) \\ &= 1 - \left(\tfrac{4}{5}\right)^8 = \frac{325\,089}{390\,625} = 0.8322 \quad \text{(to 4 d.p.)}. \end{aligned}$$

(f) Since X has a geometric distribution with parameter $p = \frac{1}{5}$, the mean number of answers required to obtain a first correct answer is $1/p = 5$.

(g) The required probability is that of providing 25 consecutive incorrect answers, which is

$$\begin{aligned} &P(X > 25) = (1 - p)^{25} \\ &= \left(\tfrac{4}{5}\right)^{25} = 0.0038 \quad \text{(to 4 d.p.)}. \end{aligned}$$

Solution 4.2

The probability of a 'triple 5' on one roll of the three dice is $\frac{1}{512}$, from Solution 3.5(b). Hence the number of rolls to obtain a 'triple 5' has a geometric distribution with parameter $p = \frac{1}{512}$, so the expected number of rolls to obtain a 'triple 5' is $1/p = 512$.

Solution 4.3

(a) Here a success is winning the hamper and a failure is not winning the hamper. We have $p = P(\text{success}) = \frac{1}{150}$ and $P(\text{failure}) = \frac{149}{150}$. In two years, there will be 104 weekly raffle draws. The probability of not winning a hamper at all during this period is

$$\left(\tfrac{149}{150}\right)^{104},$$

so the probability of winning a hamper at least once is

$$1 - \left(\tfrac{149}{150}\right)^{104} = 0.5013 \quad \text{(to 4 d.p.)},$$

which is just greater than 0.5.

(b) The purchase of n tickets in a week gives the holder of those tickets probability $n/150$ of winning the hamper. Hence at least 75 tickets are required to have a probability of at least 0.5 of winning.

(c) Since 'number of weeks to first win the hamper' has a geometric distribution with parameter $p = \frac{1}{150}$, the mean number of weeks between prizes is $1/p = 150$. On average, the customer will win a hamper about every three years.

Solution 5.1

For the first dinosaur, one lunch box has to be bought.

The probability that the second dinosaur is different in type from the first is $\frac{6}{7}$, so the number of boxes that need to be bought, on average, to obtain a dinosaur different from the first (which is a 'success' in this case) is $1/\frac{6}{7} = \frac{7}{6}$.

The total number of boxes needed, on average, to obtain the first two distinct dinosaurs is $1 + \frac{7}{6}$.

After the second dinosaur has been obtained, the probability that the next dinosaur is different in type from either of the first two is $\frac{5}{7}$, so the number of boxes that need to be bought, on average, to obtain a third dinosaur is $1/\frac{5}{7} = \frac{7}{5}$.

The total number of boxes needed, on average, to obtain the first three distinct dinosaurs is $1 + \frac{7}{6} + \frac{7}{5}$.

Continuing the argument in this way, the total number of lunch boxes which have to be purchased, on average, to obtain the whole collection is

$$1 + \tfrac{7}{6} + \tfrac{7}{5} + \tfrac{7}{4} + \tfrac{7}{3} + \tfrac{7}{2} + 7 = \tfrac{363}{20} = 18.15.$$

So approximately 18 lunch boxes would have to be bought, on average, to obtain the whole collection of dinosaurs.

It is assumed here that there are equal supplies of each type of dinosaur, so that the probability of finding any one in a particular lunch box is $\frac{1}{7}$.

Solution 5.2

(a) The first of the two particular students can choose any one of the 8 main courses. The second student then has probability $\frac{1}{8}$ of choosing the same main course. Hence the probability that both choose the same main course is $\frac{1}{8}$.

(b) Since 'at least two choose the same dish' is the same event as 'not all four choose different dishes', this question can be answered by first calculating the probability that all of the students choose different dishes.

The first student can choose any of the 8 main courses. The probability that the second student chooses a different dish is $\frac{7}{8}$. The probability that the third student chooses a dish different from either of those chosen by the first two students is $\frac{6}{8}$, and the probability that the fourth student chooses a dish different from the first three is $\frac{5}{8}$.

Hence we have

$$\begin{aligned} &P(\text{at least two the same}) \\ &= 1 - P(\text{all different}) \\ &= 1 - (\tfrac{7}{8} \times \tfrac{6}{8} \times \tfrac{5}{8}) = 0.5898 \quad \text{(to 4 d.p.).} \end{aligned}$$

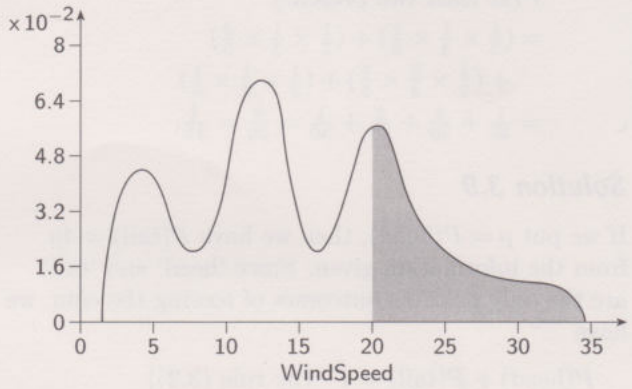
(This approach is similar to that applied to the problem of coinciding birthdays, in Chapter D1, Subsection 5.2.)

Solutions for Chapter D2

Solution 1.1

(a) The histogram has two or possibly three peaks or modes, so a normal distribution is not a suitable model for this data.

(b) A suitable curve is as follows.



(The shaded area is not part of the answer here, but is referred to in part (c) below.)

(c) Assuming that the total area under the curve sketched in part (b) is one, the shaded area represents the proportion of daily average wind speeds that are of more than 20 units in magnitude.

Solution 2.1

(a) The sample mean, $\overline{x}$, is given by

$$\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{7+5+4+5+6+9}{6} = 6.$$

(b) To find the sample standard deviation, s, the following table is useful.

x	$x-\overline{x}$	$(x-\overline{x})^2$
7	1	1
5	−1	1
4	−2	4
5	−1	1
6	0	0
9	3	9
36	0	16

Then we have

$$s = \sqrt{\frac{1}{6-1}\sum_{i=1}^{6}(x_i-\overline{x})^2} = \sqrt{\frac{16}{5}} \simeq 1.8.$$

(These results can be checked using **Summary stats...** in *OUStats*.)

Solution 2.2

(a) (i) The mean in each case is given by

$$\overline{x} = \frac{1}{5}\sum_{i=1}^{5} x_i.$$

For each of sets A and B, we obtain

$\overline{x} = 11$ marks.

(ii) The standard deviations can each be calculated from the formula

$$s = \sqrt{\frac{1}{4}\sum_{i=1}^{5}(x_i-\overline{x})^2},$$

using a table as in Solution 2.1. We obtain

$s \simeq 1.6$ marks (set A), $s \simeq 7.9$ marks (set B).

(These results can be checked using **Summary stats...** in *OUStats*.)

(b) The results show that, while the samples have the same mean, the marks in set B are considerably more variable than those in set A.

Solution 2.3

Statement (b) is false and the rest are true.

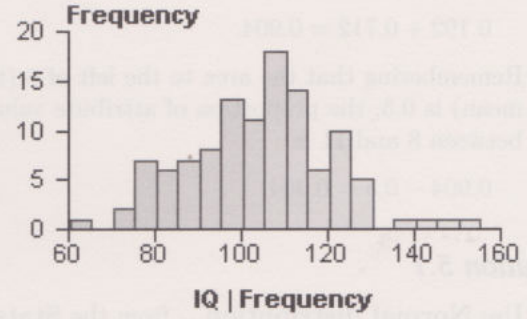
Solution 3.1

(a) The least value of IQ recorded is 61. (This can be seen either by observing the data directly, since they are presented in order of increasing IQ, or by using **Summary stats...** from the **Stats** menu and selecting 'IQ | Frequency'.)

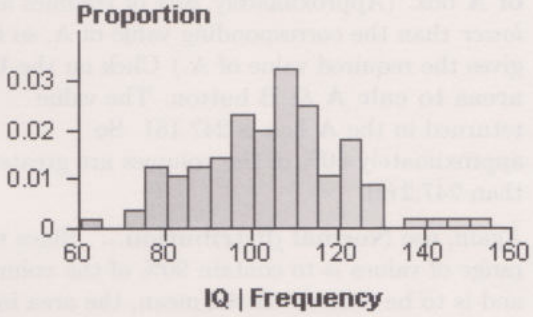
(b) Since the data are recorded to the nearest integer, the first interval starting value could be chosen as $61 - 0.5 = 60.5$, and the interval width could be 1 or some multiple of 1. Hence one could choose 60.5 as the first interval starting value, and (say) 5 for the interval width.

(c) Using **Frequency diagram...** from the **Plot** menu, and taking the first interval starting value and interval width given in part (b), the following diagrams are obtained. (You may have chosen a different interval width.)

(i) A frequency diagram:



(ii) A histogram:



(d) In a frequency diagram, the height of each bar represents the corresponding frequency, whereas in a histogram, the area of each bar is proportional to the corresponding frequency.

(Since the area under a histogram is one, it is a histogram, rather than a frequency diagram, that should be used when fitting a normal curve to data.)

Solution 3.2

The minimum value is 2.1, and the data appear to be recorded to the nearest 0.1, so the first interval starting value should be $2.1 - \frac{1}{2} \times 0.1 = 2.05$. Choosing the interval width to be 2 (you may have made a different choice) gives the histogram shown in Exercise 1.1.

Solution 4.1

Using **Summary stats...** from the **Stats** menu for this data gives

$\overline{x} \simeq 7.214$, $s \simeq 3.647$.

Hence we have the following.

(a) The sample mean is about 7.21 hours.

(b) The sample standard deviation is about 3.65 hours.

Solution 4.2

(a) The proportion for whom the attribute value is more than 6 is

$$1 - 0.192 = 0.808.$$

(b) The proportion for whom the attribute value is less than 11 is

$$0.192 + 0.712 = 0.904.$$

(c) Remembering that the area to the left of 8 (the mean) is 0.5, the proportion of attribute values between 8 and 11 is

$$0.904 - 0.5 = 0.404.$$

Solution 5.1

(a) Use **Normal distribution...** from the **Stats** menu. Enter the given mean and standard deviation, then enter 0.2 into the **Area to left of A** box. (Approximately 20% of volumes are *lower* than the corresponding value of A, so this gives the required value of A.) Click on the **Use areas to calc A & B** button. The value returned in the **A** box is 247.181. So approximately 80% of the volumes are greater than 247.2 ml.

(b) Again, use **Normal distribution...**. Since the range of values is to contain 90% of the volumes and is to be centred on the mean, the area in each tail needs to be $\frac{1}{2}(1 - 0.9) = 0.05$. Entering 0.05 into the **Area to left of A** box, and 0.9 into the **Area between A and B** box, then clicking on the **Use areas to calc A & B** button, gives 243.727 in the **A** box and 257.873 in the **B** box. Hence a range of values within which approximately 90% of the volumes lie is, in ml, $(243.7, 257.9)$.

Solution 5.2

(a) Since a normal distribution is being used as a model, about 50% of visits are completed within 32.5 minutes (the mean of the distribution).

(b) In **Normal distribution...**, with mean 32.5 and standard deviation 7.5, enter 0.9 in the **Area to left of A** box, and then click on the **Use areas to calc A & B** button. This gives a value for A of 42.1116, so 90% of visits are completed within about 42 minutes.

(c) Since the range of values is to contain 99% of the times and is to be centred on the mean, the area in each tail needs to be $\frac{1}{2}(1 - 0.99) = 0.005$. Entering 0.005 into the **Area to left of A** box, and 0.99 into the **Area between A and B** box, then clicking on the **Use areas to calc A & B** button, gives 13.1813 for A and 51.8187 for B. Hence a range of values within which approximately 99% of the times lie is, in minutes, $(13.2, 51.8)$.

Solution 5.3

(a) Use **Normal distribution...** with mean 0 and standard deviation 1 (the default). Then the required proportion of values is twice the area to the left of -1. Hence enter -1 in the **A** box, then press the **Use A & B to calc areas** button. This gives the value 0.158 655 in the **Area to left of A** box, from which the required proportion of values is twice this, about 0.317. So approximately 32% of values in a normal distribution are more than 1 standard deviation from the mean. (Another way of saying this is that approximately 68% of values are *within* 1 standard deviation of the mean.)

(b)–(e) Similarly, we obtain the following.

Standard deviations from the mean	Proportion of values
1.5	0.134
2.25	0.024
2.5	0.012
3	0.003

Solution 6.1

The underlying results needed here are summarised on pages 33 and 34 of Chapter D2. These results indicate the proportion of the attribute values in a normal distribution which lie within k standard deviations of the mean, μ (that is, the proportion that lie within the interval $(\mu - k\sigma, \mu + k\sigma)$), for certain values of k. The proportions concerned are those on which this exercise is based.

(a) According to the model, the weights of about 99.7% of all adult males in the town should lie between $\mu - 3\sigma$ and $\mu + 3\sigma$, that is, between 28.3 kg and 116.5 kg.

(b) Similarly, the weights of about 90% of the adult males should lie between

$$72.4 - 1.64 \times 14.7 \simeq 48.3\,\text{kg} \quad \text{and}$$
$$72.4 + 1.64 \times 14.7 \simeq 96.5\,\text{kg}.$$

(c) The weights of about 95% of the adult males should lie between

$$72.4 - 1.96 \times 14.7 \simeq 43.6\,\text{kg} \quad \text{and}$$
$$72.4 + 1.96 \times 14.7 \simeq 101.2\,\text{kg}.$$

(d) The weights of about 99% of the adult males should lie between

$$72.4 - 2.58 \times 14.7 \simeq 34.5\,\text{kg} \quad \text{and}$$
$$72.4 + 2.58 \times 14.7 \simeq 110.3\,\text{kg}.$$

(More accurate answers may be obtained by applying **Normal distribution...** within *OUStats*, in the manner of Solutions 5.1(b) and 5.2(c).)

Solutions for Chapter D3

Solution 1.1

(a) We apply the Central Limit Theorem. The sampling distribution of the mean for samples of size 60 has mean 251 ml and standard deviation

$$SE = \frac{2.5}{\sqrt{60}} \simeq 0.323 \text{ ml}.$$

(b) Since the sampling distribution of the mean is normal, approximately 95% of mean volumes in samples of size 60 will lie between

$$251 - 1.96 \times \frac{2.5}{\sqrt{60}} \simeq 250.4 \text{ ml} \quad \text{and}$$

$$251 + 1.96 \times \frac{2.5}{\sqrt{60}} \simeq 251.6 \text{ ml};$$

that is, they will lie in the approximate range, in ml, $(250.4, 251.6)$.

Solution 1.2

(a) The standard error of the mean (which is the standard deviation of the sampling distribution of the mean) is given by

$$SE = \frac{4.2}{\sqrt{45}} \simeq 0.626 \text{ g}.$$

(b) Approximately 95% of mean weights in samples of size 45 will lie between

$$401.6 - 1.96 \times \frac{4.2}{\sqrt{45}} \simeq 400.4 \text{ g} \quad \text{and}$$

$$401.6 + 1.96 \times \frac{4.2}{\sqrt{45}} \simeq 402.8 \text{ g};$$

that is, they will lie in the approximate range, in grams, $(400.4, 402.8)$.

Solution 2.1

(a) (i) The sample, of size $n = 50$, has mean $\overline{x} = 331.7$ and standard deviation $s = 5.75$. A 95% confidence interval for the mean volume of liquid in cans of the soft drink is therefore, in ml,

$$\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$

$$= \left(331.7 - 1.96 \times \frac{5.75}{\sqrt{50}},\ 331.7 + 1.96 \times \frac{5.75}{\sqrt{50}}\right)$$

$$\simeq (330.1, 333.3),$$

rounded to the same accuracy as the sample mean.

(ii) Since the nominal value of 330 ml is below the lower 95% confidence limit, the manufacturer can be fairly confident that the mean volume of liquid in cans of the soft drink is above the nominal value.

(b) (i) For a sample of size n, with standard deviation s, the width of the 95% confidence interval is $2 \times 1.96 \times s/\sqrt{n}$. Hence, ignoring variations in s, the width of the confidence interval is directly proportional to $1/\sqrt{n}$. Since $\frac{1}{2}(1/\sqrt{n}) = 1/\sqrt{4n}$, and n was 50 in the first sample, the second sample should have approximate size 200.

(ii) Probably not. The second sample will also be chosen at random, and it is unlikely that it will have exactly the same standard deviation s as in the first sample.

Solution 2.2

(a) The sample, of size $n = 30$, has mean $\overline{x} = 19.0$ and standard deviation $s = 10.9$. A 95% confidence interval for the mean number of words in sentences from the novel is therefore

$$\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$

$$= \left(19.0 - 1.96 \times \frac{10.9}{\sqrt{30}},\ 19.0 + 1.96 \times \frac{10.9}{\sqrt{30}}\right)$$

$$\simeq (15.1, 22.9).$$

(b) The result of part (a) is a 95% confidence interval for the *population* mean. The population here is all of the sentences in the novel. If a large number of random samples of 30 sentences are taken from this novel, then for approximately 95% of these samples, the above process will give an interval that contains the population mean. The other 5% of samples will lead to intervals that do not contain the population mean. We can be fairly confident that the mean length of sentences in the novel is between 15 and 23 words.

Solution 2.3

(a) (i) For samples of size 100, the mean of the sampling distribution of the mean is 50 hours, and the corresponding standard deviation is $7.5/\sqrt{100} = 0.75$ hours.

(ii) Approximately 95% of mean lifetimes in samples of size 100 will lie between

$$50 - 1.96 \times 0.75 \simeq 48.5 \text{ hours} \quad \text{and}$$

$$50 + 1.96 \times 0.75 \simeq 51.5 \text{ hours};$$

that is, they will lie in the approximate range, in hours, $(48.5, 51.5)$.

(iii) A range like that above can be calculated when the population mean and standard deviation are known. A confidence interval, based on the mean and standard deviation of a random sample from the population, provides an estimate (and accompanying confidence level) for an unknown population mean.

(b) (i) The sample, of size $n = 100$, has mean $\overline{x} = 49$ and standard deviation $s = 4.8$. A 95% confidence interval for the mean lifetime of batteries produced under the new procedures is therefore, in hours,

$$\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$
$$= \left(49 - 1.96 \times \frac{4.8}{\sqrt{100}},\ 49 + 1.96 \times \frac{4.8}{\sqrt{100}}\right)$$
$$\simeq (48.1, 49.9).$$

(ii) Under the new procedures there appears to be less variablity in the battery lifetimes. However, the mean lifetime seems likely to be significantly less than under the old procedures. In particular, since the 95% confidence interval does not contain 50, we can be fairly confident that the mean lifetime of all batteries manufactured under the new procedures, that is, the new population mean, will be less than the mean lifetime of 50 hours achieved under the old procedures.

Solution 2.4

(a) The sample of height difference data, of size $n = 1079$, has mean $\overline{x} = 5.15$ and standard deviation $s = 3.03$. A 95% confidence interval for the mean height difference in the population from which the sample is taken is therefore, in inches,

$$\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$
$$= \left(5.15 - 1.96 \times \frac{3.03}{\sqrt{1079}},\ 5.15 + 1.96 \times \frac{3.03}{\sqrt{1079}}\right)$$
$$\simeq (4.969, 5.331).$$

So we can be fairly confident that fathers in the population at large were, on average, approximately 5 inches taller than mothers.

(b) Since 4 is not in the confidence interval, and indeed is quite a long way below the lower confidence limit, it is very unlikely that the mean difference in height in the population was 4 inches or less, and that the observed difference of 5.15 inches was due to sampling error.

Solution 3.1

(a) Using **Confidence interval...** from the **Stats** menu, and selecting both 'SLength' and 'NLength', gives the required results. A 95% confidence interval for the mean length of Sussex weasels is given, in mm, as $(200.4, 207.1)$. A 95% confidence interval for the mean length of Northumberland weasels is given, in mm, as $(205.8, 213.8)$. (In each case, the *OUStats* output values have been rounded to 1 d.p.)

(b) For the sample of Northumberland weasels, **Confidence interval...** gives $n = 38$ for the sample size, $\overline{x} = 209.8$ for the sample mean, and $s = 12.59$ for the sample standard deviation. A 95% confidence interval for the mean length of Northumberland weasels is therefore, in mm,

$$\left(\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right)$$
$$= \left(209.8 - 1.96 \times \frac{12.59}{\sqrt{38}},\ 209.8 + 1.96 \times \frac{12.59}{\sqrt{38}}\right)$$
$$\simeq (205.8, 213.8).$$

This confidence interval agrees with that obtained directly from *OUStats* in part (a).

Solutions for Chapter D4

Solution 1.1

(a) The required quantities for the given sample, in seconds, are as follows.

median: 154.5

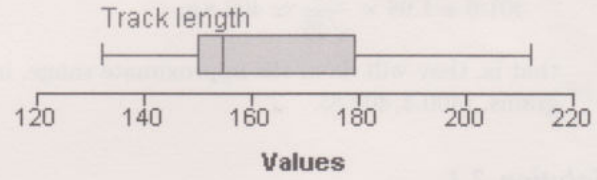
lower quartile: 150

upper quartile: 179

range: $212 - 132 = 80$

interquartile range: $179 - 150 = 29$

(b) The corresponding boxplot is as follows.



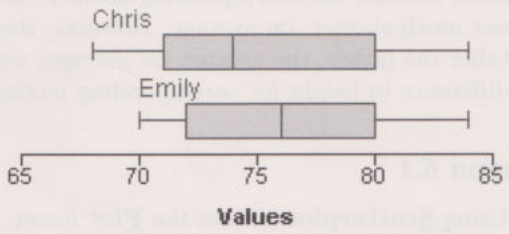
(These results can be checked using **Summary stats...** and **Boxplot...** in *OUStats*.)

Solution 1.2

(a) The required quantities for the two samples of scores are as follows.

	Chris	Emily
median	74	76
lower quartile	71	72
upper quartile	80	80
range	$84 - 68 = 16$	$84 - 70 = 14$
interquartile range	$80 - 71 = 9$	$80 - 72 = 8$

(b) The corresponding boxplots are as follows.



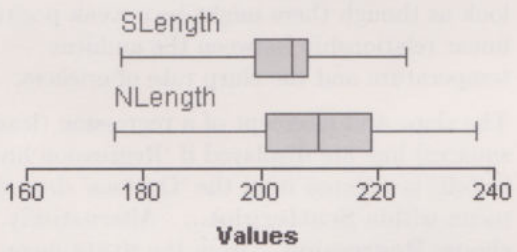
(The results for parts (a) and (b) can be checked using *OUStats*.)

(c) From the boxplots it can be seen that Emily's scores were generally a little higher than Chris's scores. Of the five key values on a boxplot, the minimum, lower quartile and median are higher for Emily, while the upper quartile and maximum are the same. Hence there is limited evidence from part (b) to support Chris's claim (since the aim in a round of golf is to obtain the lowest possible score).

(However, if the given scores correspond to each other in order, then Chris has won 6 rounds but lost 8, with 1 tied.)

Solution 2.1

(a) Using **Boxplot...** from the **Plot** menu, and selecting both 'SLength' and 'NLength', gives the boxplots shown below.



(b) In these samples, the lengths of the Northumberland weasels are generally greater than the lengths of the Sussex weasels. Four of the five key values on a boxplot are greater for the Northumberland weasels than for the Sussex weasels. However, the lengths are more variable for the Northumberland weasels, and the shortest of the Northumberland weasels is shorter than the shortest of the Sussex weasels.

Solution 3.1

We have sample sizes $n_A = n_B = 35$, sample means $\overline{x}_A = 22.3$, $\overline{x}_B = 21.4$, and sample standard deviations $s_A = 17.2$, $s_B = 13.4$.

(a) The estimated standard error is

$$ESE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$
$$= \sqrt{\frac{(17.2)^2}{35} + \frac{(13.4)^2}{35}} = 3.69 \quad \text{(to 2 d.p.)}.$$

(b) The test statistic is

$$z = \frac{\overline{x}_A - \overline{x}_B}{ESE} = \frac{22.3 - 21.4}{ESE} = 0.24 \quad \text{(to 2 d.p.)}.$$

(c) Since $-1.96 < z < 1.96$, we cannot reject the null hypothesis at the 5% significance level. The data do not provide sufficient evidence to reject the hypothesis that the mean sentence length is the same in the two books.

Solution 4.1

The null and alternative hypotheses are

$$H_0 : \mu_N = \mu_S,$$
$$H_1 : \mu_N \neq \mu_S,$$

where μ_N is the mean length of the population of male weasels in Northumberland, and μ_S is the mean length of the population of male weasels in Sussex.

Using **Two-sample z-test...** from the **Stats** menu, and choosing 'NLength' as the first variable and 'SLength' as the second, the value of the test statistic is given as $z = 2.2845$. (Choosing 'NLength' and 'SLength' in the opposite order gives $z = -2.2845$, leading to the same conclusions below.)

Since the test statistic is $z = 2.2845 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis.

We conclude that the mean length of male weasels from Northumberland is not equal to the mean length of male weasels from Sussex.

The sample mean is greater for the Northumberland weasels than for the Sussex weasels, so this suggests that the mean length of male weasels from Northumberland is greater than the mean length of male weasels from Sussex.

Solution 5.1

(a) It looks as if the data would be better fitted by a curve than by the line shown.

(b) The line is slightly too steep.

Solution 5.2

(a) The explanatory variable here is cricket chirp rate, so use equation (2). The ambient temperature predicted for a cricket chirp rate of 18.0 chirps per second is

$$y = 3.291 \times 18.0 + 25.23 = 84.468.$$

So the model predicts an ambient temperature of about 84.5°F in this case.

(b) The explanatory variable here is ambient temperature, so use equation (1). If the ambient temperature drops by 5°F, then the expected drop in the cricket chirp rate is

$$0.2119 \times 5 = 1.0595.$$

So the model predicts a reduction in the cricket chirp rate of about 1 chirp per second in this case.

Solution 5.3

(a) Here heart rate is the explanatory variable, so use equation (3). For a heart rate of 80 beats per minute, we obtain

$$y = -1.13 \times 80 + 644 = 553.6.$$

The model predicts that the corresponding reaction time will be about 554 ms.

(b) Here reaction time is the explanatory variable, so use equation (4). Since Todd's reaction time is 50 ms greater than Caragh's reaction time, Todd's predicted heart rate will be greater than Caragh's predicted heart rate by

$$-0.407 \times 50 = -20.35$$

beats per minute. Noting the minus sign, we conclude that Todd's heart rate is predicted to be about 20 beats per minute *less* than Caragh's heart rate.

(Some care is needed in communicating these results. Since Todd's reaction time is 50 ms greater than that of Caragh, he reacts more *slowly* than does Caragh. Correspondingly, his predicted heart rate is *slower* than Caragh's predicted heart rate. The relationship can be summarised in words as: the faster the reaction time, the faster the expected heart rate.)

Solution 5.4

(a) The estimated mean heights of the mothers (in inches) are the values of MHt below. (Note that it is inappropriate to give these estimates to an accuracy of more than 1 decimal place, because the data heights were measured only to the nearest $\frac{1}{4}$ inch.) In each case, we have

$$\text{MHt} = 44.98 + 0.2595 \times \text{FHt}.$$

(i) FHt = 62 : MHt $\simeq$ 61.1

(ii) FHt = 66 : MHt $\simeq$ 62.1

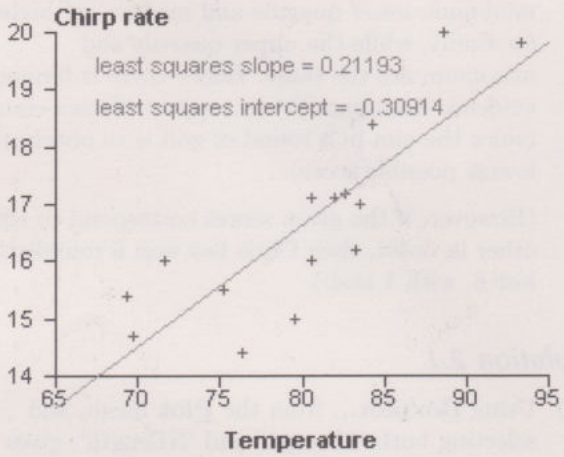
(iii) FHt = 70 : MHt $\simeq$ 63.1

(iv) FHt = 73 : MHt $\simeq$ 63.9

(b) These results indicate that when compared with short fathers, the corresponding mothers were not much shorter, on average. However, the taller the father, the greater, on average, was the difference in height for corresponding mothers.

Solution 6.1

(a) Using **Scatterplot...** from the **Plot** menu, with ambient temperature as the x variable and chirp rate as the y variable, gives the scatterplot below. The regression line is added by opening the 'Options' drop-down menu and clicking on 'Regression line on/off'.



(b) There is a lot of scatter in the plot, but it does look as though there might be a weak positive linear relationship between the ambient temperature and the chirp rate of crickets.

(c) The slope and intercept of a regression (least squares) line are displayed if 'Regression line on/off' is selected from the 'Options' drop-down menu within **Scatterplot...**. Alternatively, choose **Regression...** from the **Stats** menu.

(i) Selecting ambient temperature as the x variable and chirp rate as the y variable gives the equation of the regression line of y on x as

$$y = -0.3091 + 0.2119x.$$

(This is the line which appears on the scatterplot in part (a) above.)

(ii) Selecting chirp rate as the x variable and ambient temperature as the y variable gives the equation of the regression line of y on x as

$$y = 25.23 + 3.291x.$$

(These two equations of regression lines are quoted in Exercise 5.2.)

Printed in the United Kingdom by
Lithmark Limited